Year 4 — Numerical Linear Algebra

Based on lectures by Prof. Yuji Nakatsukasa
Notes taken by James Arthur

Michaelmas 2022

These notes are not endorsed by the lecturers, and I have modified them (often significantly) after lectures. They are nowhere near accurate representations of what was actually lectured, and in particular, all errors are almost surely mine (especially the typos!).

Contents

1 Introduction

The goal of this course is to solve and understand problems using (usually non-square) matrices. The course will cover,

1. SVD. Singular Value Decomposition, see Linear Algebra (Year 2). $A = U\Sigma V^T$.
2. Linear systems $Ax = b$ and eigenvalue problems $Ax = \lambda x$. Least square problems, $\min_x \|Ax - b\|_2$.

There are going to be three classes of approaches,

- Direct Methods, classical approach. Not useful for very big matrices.
- Iterative solvers, work even if matrix sizes are larger.
- Randomised Algorithms, use some sense of randomisation in order to get an algorithm that works with high probability better than direct for very large matrices.

1.1 Why do we care?

When we want to solve a non-linear equation a linear system pops out. Motivation: minimising a function, $\min_x f(x)$ where $f : \mathbb{R}^n \to \mathbb{R}$, where we let n be large. One powerful way to find a minimum is to find a stationary point. So we find,

$$\nabla f = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \dots \\ \frac{\partial f}{\partial x_n} \end{pmatrix} = 0$$

if we have a nice (convex) function, then if we have a stationary point, we have a minimiser. This boils down to, letting $F = \nabla f : \mathbb{R}^n \to \mathbb{R}^n$. We need to find zeros of this non-linear problem. Hence we now have a root-finding problem, and so we can use Newton's Method.

$$x_{\text{new}} = x_{\text{old}} - \mathcal{J}^{-1} F(x_{\text{old}})$$

and we see,

$$\mathcal{J}_{ij} = \frac{\partial F_i}{\partial x_j}$$

and then this is the hessian of f. This is then a linear system,

$$\mathcal{J} \Delta x = F(x_{\text{old}}).$$

NB! Here $A^* = \overline{A}$ (instead of $A^* = \overline{A}^T$) We are going to stay under real matrices because there are only two cases where the difference matters. Further, $m \geq n$ for most matrices in this course. For orthonormal matrices,

$$A^T A = I_n \qquad AA^T \neq I_m$$

1.2 Norms

We need norms to quantify how big matrices are. These help us when approximating matrices. Given some $\mathbf{x} \in \mathbb{R}^n$, we have

$$\|x\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

and 1-norm,

$$\|x\|_1 = |x_1| + \cdots + |x_n|$$

and the p-norm,

$$\|x\|_p = (|x_1|^p + \cdots + |x_p|^p)^{\frac{1}{p}}$$

and finally the ∞-norm,

$$\|x\|_\infty = \max_i |x_i|$$

Definition 1.1 (Norm)**.** A norm satisfies three axioms,

- $\|\alpha x\| = |\alpha| \, \|x\|$
- $\|x\| \geq 0$ and $\|x\| = 0 \iff x = 0$
- $\|x + y\| \leq \|x\| + \|y\|$

Lemma 1.2 (Holder's Inequality)**.** For $p > q$,

$$\|x\|_p \leq \|x\|_q$$

Definition 1.3 (Unitarily Invariant)**.** If A is orthonormal, then $\|Ax\|_2 = \|x\|_2$.

Lemma 1.4 (Cauchy-Schwartz)**.** For any $x, y \in \mathbb{R}^n$,

$$|x^T y| \leq \|x\|_2 \, \|x\|_2$$

Proof. For any scalar c, $\|x - cy\|_2^2 = \|x\|_2^2 - 2cc^T y + \|y\|_2^2$. Now we complete the square and we get,

$$\begin{aligned}
\|x - cy\|_2^2 &= \|x\|_2^2 - 2cc^T y + \|y\|_2^2 \\
&= \|y\|_2^2 \left(c - \frac{x^T y}{\|y\|_2} \right)^2 + \|x\|_2^2 - \frac{(x^T y)^2}{\|y\|_2^2}
\end{aligned}$$

and so we now minimise by letting $c = \frac{x^T y}{\|y\|_2}$ and so we get Cauchy Schwartz. $\square$

Now for matrix norms. We have the p-norm,

$$\|A\|_p = \sup_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p} = \max_{\|x\|_p = 1} \frac{\|Ax\|_p}{\|x\|_p}.$$

The most important case is when $p = 2$, this is the spectrum norm. This is,

$$\|A\|_2 = \max_{\|x\|_2 = 1} \|Ax\|_2$$

Exercise. Show,

$$\begin{aligned}
\|A\|_1 &= \text{maximum column sum} \\
\|A\|_\infty &= \text{maximum row sum}
\end{aligned}$$

Definition 1.5 (Frobenius Norm)**.**

$$\|A\|_F = \sqrt{\sum_i \sum_j |A_{ij}|^2} = \sqrt{\mathbf{A}^T \mathbf{A}} = \sqrt{\operatorname{tr}(A^T A)}$$

where

$$\mathbf{A} = \begin{pmatrix} A_{11} \\ \vdots \\ A_{1n} \\ A_{21} \dots \end{pmatrix}$$

Definition 1.6 (Trace Norm)**.**

$$\|A\|_* = \sum_{i=1}^{\min(m,n)} \sigma_i(A)$$

For most p-norms,

$$\|AB\|_p \leq \|A\|_p \|B\|_p$$

and for the Frobenius norm,

$$\begin{aligned}\|AB\|_F &\leq \|A\|_F \|B\|_F \\ &\leq \|A\|_2 \|B\|_F\end{aligned}$$

this is the subordinate property. Where this goes wrong is,

$$\|A\|_\infty = \max_{i,j} |A_{ij}|$$

Here is a result on subspaces,

Lemma 1.7. $\mathcal{S}_1 = \operatorname{span}(V_1)$ and $\mathcal{S}_2 = \operatorname{span}(V_2)$ where $V_1 \in \mathbb{R}^{n \times d_1}$ and $V_2 \in \mathbb{R}^{n \times d_2}$, with $d_1 + d_2 > n$. Then $\exists x \in 0 \in \mathcal{S}_1 \cap \mathcal{S}_2$.

2 SVD

We already know of the symmetric eigenvalue decomposition $A = V\Lambda^T V$ for a symmetric $A \in \mathbb{R}^{n\times n}$ where $V^T V = I_n$ and $\Lambda = \mathrm{diag}(\lambda_1, \dots \lambda_n)$. Now we learn about the Singular Value Decomposition (SVD). This is for any $A \in \mathbb{R}^{m\times n}$ for $m \geq n$. Here $U^T U = V^T V = I_n$, $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$ where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$. We now want to prove this always exists,

Proof. Take some $A^T A$ (Gram Matrix) symmetric positive semi-definite. That is, all the eigenvalues are nonnegative. We then have an eigenvalue decomposition,

$$A^T A = V \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} V^T$$

Now we let $B = AV$. Then,

$$\begin{aligned} B^T B &= V^T A^T A V \\ &= \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{pmatrix} = \Sigma^2 \end{aligned}$$

Suppose $\lambda_n > 0$. Then let,

$$U = B \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & \\ & \ddots & \\ & & \lambda_n^{\frac{1}{2}} \end{pmatrix}$$

and $U^T U = \Sigma^{-1} B^T B \Sigma^{-1} = I_n$. Now $A = BV^T$ and $B = U\Sigma$ and so $A = U\Sigma V^T$. Now consider when $\lambda_{r+1} = 0$. We have,

$$\Sigma = \begin{pmatrix} \lambda_1 & & & & & \\ & \ddots & & & & \\ & & \lambda_r & & & \\ & & & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{pmatrix}$$

and now we can do something sensible. We do,

$$\begin{pmatrix} U_r & 0 \end{pmatrix} = B \begin{pmatrix} \lambda_1^{\frac{1}{2}} & & & & & \\ & \ddots & & & & \\ & & \lambda_r^{\frac{1}{2}} & & & \\ & & & 1 & & \\ & & & & \ddots & \\ & & & & & 1 \end{pmatrix}$$

instead of 0's. We still have $A = BV^T$, but,

$$U = BV^T = \begin{pmatrix} U_r & 0 \end{pmatrix} \begin{pmatrix} \Sigma_r & \\ & I \end{pmatrix} \begin{pmatrix} V_r^T \\ V_\perp^T \end{pmatrix} = U\Sigma_r V^T$$

This is the economical SVD. $\square$

We also have the full SVD where, $A = \begin{pmatrix} U & U_\perp \end{pmatrix} \begin{pmatrix} \Sigma \\ 0 \end{pmatrix} V^T$ where $U \in \mathbb{R}^{m\times m}$ orthogonal.

From the SVD we get,

- rank r of $A \in \mathbb{R}^{m\times n}$, number of nonzero singular values $\sigma_i(A)$. We can always write $A = \sum_{i=1}^{\text{rank}(A)} \sigma_i u_i v_i^T$,
- Column Space, span of $U = [u_1, \dots, u_r]$,
- row spane, row span of $v_1^T, \dots, v_r^T$,
- null space, $v_{r+1}, \dots, v_n$.

We note that the SVD can be written in the form of an outer product, $A = U\Sigma V^T = \sum_{i=1}^{k} \sigma_i U_i V_i^T$. We also note that $Av_i = \sigma_i u_i$. Let's prove this,

$$\begin{aligned} Av_i &= \left(\sum_{j=0}^{n} \sigma_j u_j v_i^T\right) v_i \\ &= \sigma_i u_i \end{aligned}$$

and similarly, $u_i^T A = \sigma_i v_i^T$. Further we can show if $A = U\Sigma V^T$ we can say,

$$A^T A = V\Sigma^2 V^T \tag{1}$$

or

$$AA^T = U\Sigma^2 U^T \tag{2}$$

Further for any 1 such that $A = \tilde{U}\Sigma V^T$ and $A = U\Sigma\tilde{V}^T$ then **some result**.
We call a triplet a singular triple if $Av_i = \sigma_i u_i$ and $u_i^T A = \sigma_i v_i^T$ both hold. We now ask whether the SVD is unique. We can see,

$$\begin{aligned} A &= U\Sigma V^T \\ &= USS^{-1}\Sigma SS^{-1}V^T \\ &= (US)(S\Sigma S)(SV^T) \end{aligned}$$

where we just let S be diagonal with ± 1. We see that the grouped terms retain the structure that we need to have a new SVD. Further this SVD is different, but Σ is the same ($S = S^{-1}$ and Σ is diagonal). When $\sigma_i = \sigma_j$, we have larger degree og freedom, i.e. $\sigma_1 = \sigma_2$. We can write the first two diagonals as $\sigma_1 QQ^T$. We then have,

$$A = U \begin{pmatrix} Q & \\ & I_{n-2} \end{pmatrix} \Sigma \begin{pmatrix} Q^T & \\ & I_{n-2} \end{pmatrix} V^T$$

If A is orthogonal, we can say the following are SVD's of A,

$$A = AII = IIA = (AQ)IQ^T.$$

Lemma 2.1.

$$\|A\|_2 = \sigma_1(A)$$

Proof. Use SVD,

$$\begin{aligned} \|Ax\|_2 &= \left\|U\Sigma V^T x\right\|_2 \\ &= \left\|\Sigma V^T x\right\| \\ &= \|\Sigma y\|_2 \end{aligned}$$

$$
\begin{aligned}
&= \sqrt{\sum_{i=1}^{n} \sigma_i^2 y_i^2} \\
&\leq \sqrt{\sum_{i=1}^{n} \sigma_1^2 y_i^2} = \sigma_1 \left\| y \right\|_2^2 = \sigma_1
\end{aligned}
$$

□

2.1 Applications

2.1.1 Low-rank approximations

Given some $A \in \mathbb{R}^{m \times n}$, we want to find some sort of A_r such that $A =\approx A_r = U_r \Sigma_r V_r^T$. You may ask why?

- Let A be stupidly large, then this saves storage.
- Matrix multiplication is of order $\mathcal{O}(mn)$, but in terms of this approximation we have something of the order of $\mathcal{O}((m+n)r)$, which is a massive saving.

Low-rank Matrices When talking about low-rank matrices, we say some $\text{rank}(B) \leq r$. If this is true we can write it as, $B = xy^T$. Let's prove this,

Proof. We know $B = U \Sigma V^T$, then we can truncate $\Sigma = \text{diag}(\sigma_1, \dots, \sigma_r)$ and then let $U_r \Sigma = X$ and $Y = V$. Conversely, if $B = XY^T$, then $B = U_X \Sigma_X V_X^T$ then $Y = U_Y \Sigma_Y V_Y^T$, then we look at $XY^T = U_X \Sigma) X V_X^T V_Y \Sigma_Y U_Y^T$. Now we SVD AGAIN!!!!! $V = U_X \tilde{U} \tilde{\Sigma} \tilde{V}^T U_Y^T$ AND AGAIN! $\tilde{\tilde{U}} \tilde{\tilde{\Sigma}} \tilde{\tilde{V}}^T$. Then we can say that V can have at most r positive singular values. □

Now we want to find B such that $\text{rank}(B) \leq r$. So we want to minimise $\left\| A - B \right\|_2$ given A. We see $\left\| A - B \right\|_2 \leq \left\| A - C \right\|_2$ for all $\text{rank}\, C \leq r$. That is, we want to solve,

$$
A = U \Sigma V^T = \sum_{n=1}^{n} s_i u_i v_i
$$

and then truncating B,

$$
B = U_r \Sigma_r V_r^T = \sum_{i=0}^{r} \sigma_i u_i v_i^T
$$

w